Semi-parametric analysis of multi-rater data

Simon Rogers
Department of Computing Science
University of Glasgow
G12 8QQ
srogers@dcs.gla.ac.uk

Mark Girolami
Department of Computing Science
University of Glasgow
G12 8QQ
girolami@dcs.gla.ac.uk

Tamara Polajnar
Department of Computing Science
University of Glasgow
G12 8QQ
tamara@dcs.gla.ac.uk

November 13, 2008

Abstract

Datasets that are subjectively labeled by a number of experts are becoming more common in tasks such as biological text annotation where class definitions are necessarily somewhat subjective. Standard classification and regression models are not suited to multiple labels and typically a pre-processing step (normally assigning the majority class) is performed. We propose Bayesian models for classification and ordinal regression that naturally incorporate multiple expert opinions in defining predictive distributions. The models make use of Gaussian process priors, resulting in great flexibility and particular suitability to text based problems where the number of covariates can be far greater than the number of data instances. We show that using all labels rather than just the majority improves performance on a recent biological dataset.

1 Introduction

In the traditional predictive modeling setting, one is presented with a set of M-vector of covariates, $\mathbf{x}_n \in \Re^M, n = 1 \dots N$, corresponding to measurements of some set of features, and associated scalar targets, t_n. These targets may be real valued in the case of regression, one of K possible heterogeneous categories in classification or one of K ordered categories in ordinal regression. In classification and ordinal regression, these target values are generally assumed to be noise free and clearly defined. However, applications do exist where a single *true* labeling is not available and we must instead work from labels provided by several experts, each of which have their own level of variability and interpretation of the class definitions (classification) or thresholds (ordinal regression). A recent example within the classification framework is the 2007 Computational Medicine Center (CMC) Medical Natural Language Processing (NLP) challenge. The task involved automatically assigning codes to medical reports and the training and test corpora were labeled independently by three separate companies[1], with considerable expert disagreement. A classic ordinal regression example, discussed in [13], consists of student essays, each of which has been graded on an ordered scale by several examiners. Faced with such data, there

[1]Described in depth on the CMC website http://www.computationalmedicine.org/challenge/index.php

are several questions we may wish to answer. For example, in the ordinal regression case, what is the most likely ordering of the essays given the marks from the various examiners? In classification, we may wish to make predictions that more accurately encapsulate the uncertainty in the domain - disagreement between experts may not just be due to mistakes but rather due to unavoidably ambiguous class definitions. This latter example is becoming particularly important in the field of biological text annotation where disagreements between annotators are commonplace, see [18] for an interesting discussion.

Whilst a combination strategy such as taking the majority will be successful in a classification application where the differences arise through random errors, we argue that in some examples disagreements arise because of perfectly natural variations in how different experts interpret subjective class definitions. As such, keeping only a majority labeling results in a loss of information - for example, if we have three experts, two of whom label a particular instance as belonging to class A while the other assigns it to class B, by taking a majority we are saying that this instance can teach us nothing of class B, which is unlikely to be the case. This becomes more important as the number of experts increases - it seems particularly foolish to retain only majority labels when the number of experts making up the minority increases. There is also no guarantee that there will be a clear majority. Datasets with these characteristics are beginning to appear, motivating the development of predictive methods that can sensibly incorporate diverse expert opinions. A recent study, [18] propose a set of five qualitative dimensions that can be used to annotate biological text, which, when used by 12 experts, result in a 70-80% inter-annotater agreement. The authors are pleased with this figure and believe that it justifies their choice of dimensions. However, it still corresponds to a high level of disagreement and serves to show that whilst the choice and definition of the classes is important, so is the development of methods that incorporate diversity in opinion. Additionally, [5] emphasises that discussion of inter-annotator agreement may be crucial for wider usage of a corpus and the success of any prediction systems on which it is based. Although many corpora now include the annotator disagreement statistics, few release the original annotations. Oahur results indicate that knowledge encoded in the multiple annotations may be crucial for predictive systems. Some differences in annotation can stem from valid ambiguities in the data, and the removal of conflicting annotations excludes potentially important information. This is true of many language related tasks; for example, [17] shows that in co-reference resolution some disagreement arises because a reference (he, she, it, etc.) does not always clearly refer to a single named entity.

As described in [15], the problem of classification in the presence of inconsistent labelings has appeared in several forms in the statistics and machine learning literature. Most frequently in medical statistics where a diagnosis or a decision on the appropriateness of a particular treatment is often based on the opinions of several experts (for example [16]). A second example is automatic object discovery in images (for example [15]). [7] proposed an expectation maximisation procedure to calculate a posterior distribution over some true latent class, conditioned on the subjective labeling, a technique that has been used more recently by [15]. As a pre-processing step, this seems more sensible than taking the majority but it relies completely on the label information and does not use covariate information.

As discussed in [15], subjectively labeled data present two main problems. Firstly, how do we train inference methods in the presence of multiple annotations? Secondly, how do we evaluate the performance of these (particularly classification) methods when the ground 'truth' consists of several, possibly conflicting opinions? In this work, we propose models that can overcome the first problem for both classification and ordinal regression. These approaches are built around Gaussian Processes (GPs) whose semi-parametric[2] nature makes them particularly powerful in domains where the feature space is very large, for example text classification and annotation (other non-parametric classifiers, particularly sup-

[2]Here we consider semi-parametric models as those where the number of parameters increases linearly with the number of data instances and not with the number of covariates.

port vector machines have been used extensively in such domains, for example [3]). The proposed solutions scale linearly with both the number of classes/categories and the number of experts. Whilst we have only briefly addressed the second problem, that of evaluation, we believe that the development of evaluative methods is an interesting area for future development.

The remainder of the paper is structured as follows. In section 2we present the ordinal regression and classification models. In section 3 we provide illustrative examples on synthetic data for both models and in sections 4 and 5 we show the performance of the algorithms on synthetic and real data respectively. Finally, in section 6 we draw some conclusions.

2 The Model

2.1 Ordinal Regression

Figure 1 about here

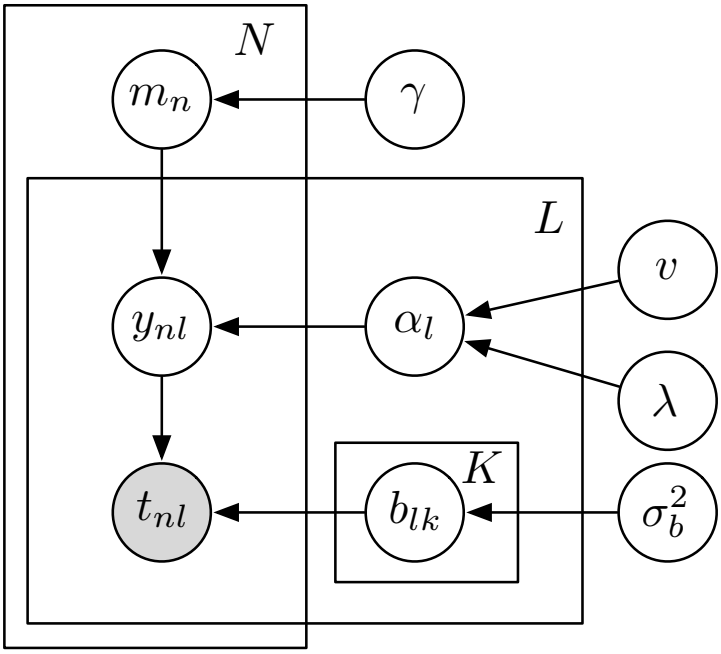


Figure 1: Plate diagram for ordinal regression model

Consider the following problem. We are given a set of N data points, $\mathbf{x}_n \in \Re^M$, and corresponding scalar targets $t_n \in \{1 \dots K\}$ where there exists an ordering amongst the classes - for example, they may be scores awarded for a school assignment. A traditional way of modeling this is to assume that the labels were produced by applying a set of unobserved thresholds (b_k) to a continuous, latent function $m(\mathbf{x}_n)$. For example, if there are K ordered classes there are $K-1$ free threshold parameters given by $b_0 = -\infty, b_1, b_{K-1}, b_K = \infty$. In the simple case of $K = 2$, this leaves only one free parameter (b_1) which is often set to 0. The inference task is thus to compute the posterior distribution over $\mathbf{m} = [m(\mathbf{x}_1), \dots, m(\mathbf{x}_N)]^T$ and $\mathbf{b} = [b_0, \dots, b_K]^T$ conditioned on the observed data and labels. Armed with these distributions we can make predictions regarding the latent variable, $\mathbf{m}$, and category for a

new data point, $\mathbf{x}_{new}$.

Following [4], we can define the likelihood function of the nth target conditioned on the corresponding value of the latent function, $m_n = f(\mathbf{m}_n)$

$$p(t_n|m_n) = \begin{cases} 1 & \text{if } b_{t_n-1} < m_n \leq b_{t_n} \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

i.e., if $t_n = 2$, m_n must be between b_1 and b_2. However, this assumes that there is no noise present in either the data or labels. Adding Gaussian noise, with mean zero and precision α, results in the following likelihood

$$p(t_n|m_n, \alpha, \mathbf{b}) = \int \mathbf{1}(b_{t_n-1} < y_n \leq b_{t_n})\mathcal{N}(y_n|m_n, \alpha^{-1})dy_n \tag{2}$$

where $\mathbf{1}(expr)$ is 1 if $expr$ is true and 0 otherwise. This can be expressed as

$$p(t_n|m_n, \alpha, \mathbf{b}) = \Phi(\alpha^{\frac{1}{2}}(b_{t_n} - m_n)) - \Phi(\alpha^{\frac{1}{2}}(b_{t_n-1} - m_n)), \tag{3}$$

the difference between two standardised normal cdf functions, $\Phi(z) = \int_{-\infty}^{z} e^{-z^2/2}dz$. There are two problems with this likelihood. Firstly, it is un-identifiable - multiplying all values of b_{t_n} and m_n by some constant a^{-1} and multiplying α by a^2 will leave the likelihood unchanged. Similarly, adding a constant value to all values of b_{t_n} and m_n does not change the likelihood. We will discuss this problem in the context of the multi-rater model later in this section. Secondly, this form of likelihood makes computation of the posterior distributions intractable and it is necessary to resort to Metropolis-Hasting sampling or approximate strategies, for example Expectation Propagation (EP) [4]. Alternatively, it is possible to remove the marginalisation from equation 2 and treat y as an additional parameter in the model. This 'auxiliary variable' method has been proposed in a linear regression setting by [2] and in a Gaussian process framework by [9]. Whilst increasing the number of variables in the model, it makes posterior inference straightforward as, assuming conjugate priors, one can easily obtain the conditional distributions necessary to implement a Gibbs sampling scheme.

Particularly, if we define a Gamma prior[3] on α (with hyper-paramters v and λ) and the priors on b_k and $\mathbf{m}$ as $p(b_k)$ and $p(\mathbf{m})$ (neglecting any hyper-parameters at this point), the Gibbs sampling distributions are

$$\begin{aligned}
p(y_n|m_n, t_n, \alpha, \mathbf{b}) &\propto \mathbf{1}(b_{t_n-1} < y_n \leq b_{t_n})\mathcal{N}(y_n|m_n, \alpha^{-1}) && (4)\\
p(b_k|y_1, \ldots, y_N) &\propto \mathbf{1}\left(\max_{n,t_n=k} y_n \leq b_k < \min_{n,t_n=k+1} y_n\right)p(b_k) && (5)\\
p(\mathbf{m}|\mathbf{y}, \alpha) &\propto \mathcal{N}(\mathbf{y}|\mathbf{m}, \alpha^{-1}\mathbf{I}_N)p(\mathbf{m}) && (6)\\
p(\alpha|\mathbf{y}, \mathbf{m}, v, \lambda) &= \mathcal{G}\left(v + \frac{N}{2}, \lambda + \frac{1}{2}(\mathbf{m} - \mathbf{y})^T(\mathbf{m} - \mathbf{y})\right) && (7)
\end{aligned}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ and $\mathbf{I}_N$ is an $N \times N$ identity matrix. The conditionals for y_n and b_k are equivalent to those described in [2].

We now extend this model to the case of L expert raters. A graphical representation of the model can be seen in figure 1. We assume that each expert, $l = 1 \ldots L$, has their own set of thresholds $b_{l0} < \cdots < b_{lK}$ and their own noisy view of the underlying function m, through their own auxiliary parameter, $\mathbf{y}_l$, the

[3] $\mathcal{G}(\alpha|v, \lambda) = \frac{\lambda^v}{\Gamma v}(\alpha)^{v-1}e^{-\lambda\alpha}$

nth component of which is *a-priori* distributed as a Gaussian with mean m_n and precision α_l. These assumptions produce the model proposed in [13]. The sampling distributions become

$$p(y_{nl}|m_n, t_{nl}, \alpha_l) \propto \mathbf{1}(b_{l,t_{nl}-1} < y_{nl} \leq b_{l,t_{nl}})\mathcal{N}(y_{nl}|m_n, \alpha_l^{-1}) \quad (8)$$

$$p(b_{lk}|\mathbf{Y}_{\cdot l}) \propto \mathbf{1}\left(\max_{n,t_{nl}=k} y_{nl} \leq b_{lk} < \min_{n,t_{nl}=k+1} y_{nl}\right) p(b_{lk}) \quad (9)$$

$$p(\mathbf{m}|\mathbf{Y}, \boldsymbol{\alpha}) \propto \prod_{l=1}^{L} \mathcal{N}(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l^{-1}\mathbf{I}_N)p(\mathbf{m}) \quad (10)$$

$$p(\alpha_l|\mathbf{Y}_{\cdot l}, \mathbf{m}, v, \lambda) = \mathcal{G}\left(v + \frac{N}{2}, \lambda + \frac{1}{2}(\mathbf{m} - \mathbf{Y}_{\cdot l})^T(\mathbf{m} - \mathbf{Y}_{\cdot l})\right), \quad (11)$$

where we have denoted the lth column of the $N \times L$ matrix $\mathbf{Y}$ as $\mathbf{Y}_{\cdot l}$ and the nth row as $\mathbf{Y}_{n\cdot}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^T$. Again, the conditionals for y_{nl} and b_{lk} are equivalent to those in [2].

2.1.1 Identifiability

We now turn to the problem of identifiability, discussed at some length for this form of likelihood in [13] and [12]. As already mentioned, simple multiplicative and additive transformations of the parameters leads to the same likelihood value. There are two schemes that could be adopted to overcome this problem - fixing the values of selected parameters to force the likelihood to be identifiable (often described as **strong identifiability**) or placing proper priors on the parameters that are sufficiently strong to overcome parameter coupling (**weak identifiability**). For the former case, identifiability in single rater ordinal models can be ensured by fixing one of the threshold values (see, for example [6]). In the multi-rater model with rater-specific precisions, we have two degrees of freedom and thus require two constraints. The most obvious way to do this is to fix one threshold (say b_{11}) and one precision (α_1). This effectively fixes the latent scale of $\mathbf{m}$ that cannot then be changed by other raters. Alternatively, one can follow [11] by leaving the precisions free and fixing two thresholds. Whilst fixing parameters in this manner does ensure identifiability, it comes at the cost of losing some model flexibility. In particular we lose any concept of posterior variability in these parameters.

To overcome this limitation, weak identifiability is adopted in, for example [11, 12]. This involves the use of appropriate prior distributions on all parameters that are strong enough to make the posterior identifiable even though the likelihood is not. This approach must be used with caution as ensuring that the prior distributions are strong enough is clearly not trivial. Particularly, as the quantity of data is increased, the prior effect will lessen suggesting that there is no single prior that is suitable for all datasets. For this reason, if one is wary about fixing the values of particular parameters and opts for weak identifiability, it is recommended that comparisons are made with the strong case to ensure that the sampling scheme is converging in a satisfactory manner. In the datasets examined in this work, comparisons between the two approaches have shown that weak identifiability is satisfactory with reasonably standard priors on α and b although in some cases convergence is improved by opting for strong identifiability (i.e., less thinning of the output samples is required). However, it should be noted that these datasets are reasonably small and stronger priors would most likely be needed in the presence of more data. The priors that we have found suitable are Gamma for the rater-specific precisions and normals (with appropriate order constraints) for the thresholds.

2.1.2 Priors on latent functions

Whether opting for strong or weak identifiability, the prior specification for the latent function is important. In the original multirater model of [13], no covariate information was included and the prior on

m_n was $\mathcal{N}(m_n|0,1)$ providing weak identifiability. The model was then extended to include covariate information in a linear regression framework

$$m_n = \boldsymbol{\beta}^T \mathbf{x}_n. \tag{12}$$

Unfortunately, this does not guarantee weak identifiability unless a suitable prior is placed on $\boldsymbol{\beta}$ which may not be possible [12]. The authors overcame this by redefining the regression model such that rather than defining $p(\mathbf{m}|\boldsymbol{\beta})$, as would be typical for a regression model, the model defines $p(\boldsymbol{\beta}|\mathbf{m})$. In this manner, a proper prior could still be defined for $\mathbf{m}$ regardless of whether or not it was possible for $\boldsymbol{\beta}$.

As an alternative, we propose placing a GP prior directly onto the latent function m. This has several major benefits. Firstly, we are able to incorporate covariate information through a suitable covariance function. Secondly, we are not restricted to a particular functional form (e.g. linear) chosen *a-priori* - [13, 12] discuss the limitations of the linear model. Thirdly, semi-parametric methods such as these are known to perform well when the number of covariates is greater than the number of data examples (e.g. [3]), a regime in which a linear model would require a high level of regularization. Finally, the covariance functions exist for all manner of data types (reals, integers, strings etc.) making the GP suitable in many domains in which other methods are not applicable. This is particularly interesting for text analysis - with the GP prior we are able to investigating more sophisticated representations.

2.2 A Gaussian process prior

Defining a GP prior on $\mathbf{m}$ with zero mean function and $N \times N$ covariance matrix $\mathbf{C}$ (created by evaluating a suitable covariance function $C(\mathbf{x}_i, \mathbf{x}_k)$ for each pair of data points) and assuming a Gaussian prior on the thresholds, subject to the necessary order constraints (i.e., a particular threshold for one rater b_{lk} conditioned on the other thresholds for that rater $b_{l0}, \ldots, b_{l,k-1}, b_{l,k+1}, \ldots, b_{lK}$ is a Gaussian truncated between $b_{l,k-1}$ and $b_{l,k+1}$), the full prior model specification is given as

$$\begin{aligned}
\mathbf{m}|\mathbf{x}_1, \ldots, \mathbf{x}_N, \gamma &\sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \\
\alpha_l|v, \lambda &\sim \mathcal{G}(v, \lambda) \\
y_{nl}|m_n, \alpha_l &\sim \mathcal{N}(m_n, \alpha_l^{-1}) \\
b_{lk}|b_{l,k-1}, b_{l,k+1}, \sigma_b^2 &\sim \mathbf{1}(b_{l,k-1} < b_{lk} < b_{l,k+1})\mathcal{N}(0, \sigma_b^2) \\
t_{nl} = k|y_{nl}, b_{l,k-1}, b_{lk} &\sim \mathbf{1}(b_{l,k-1} < y_{nl} \leq b_{lk})
\end{aligned}$$

The conditional distributions for $\mathbf{m}$ and b_{lk} required in the Gibbs sampler are given by

$$p(\mathbf{m}|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{C}) = \mathcal{N}\left(\mathbf{m}|\boldsymbol{\Sigma}\sum_{l=1}^{L}\alpha_l \mathbf{Y}_{\cdot l}, \boldsymbol{\Sigma}\right), \quad \boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \mathbf{I}_N \sum_{l=1}^{L}\alpha_l\right)^{-1}$$

$$p(b_{lk}|\mathbf{Y}_{\cdot l}, \sigma_b^2) \propto \mathbf{1}\left(\max_{n, t_{nl}=k} y_{nl} \leq b_{lk} < \min_{n, t_{nl}=k+1} y_{nl}\right)\mathcal{N}(b_{lk}|0, \sigma_b^2).$$

It is worth noting that in the extreme case that $\mathbf{C} = \mathbf{I}_N$, we have no covariate information and a model equivalent to the initial model of [13] and [12]. We will use this model for comparison in our experimental section.

More detailed derivations of the conditional distributions and a description of the sampling scheme can be found in appendix A.

2.3 Multi-class Classification

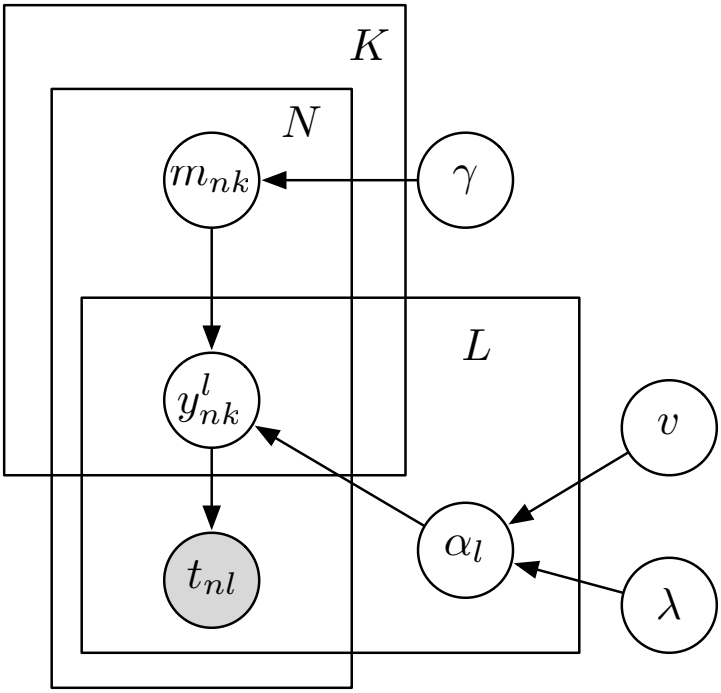


Figure 2 about here

Figure 2: Plate diagram for multi-class classification model

We now turn our attention to the classification model. In [9] an efficient, multi-class GP classifier was introduced. For an introduction to classification with GPs, see for example [19]. As in the ordinal regression scheme discussed above, posterior inference within this model is made more tractable through the addition of a set of auxiliary variables. It is perhaps easiest to think of the classifier in a generative sense. With one expert, a dataset of N examples spanning K classes is generated as follows: For each class $k = 1 \dots K$, realisations of the GP, at the location of the N datapoints, m_{nk}, are drawn from the GP prior, with zero mean and user defined covariance function. As for the ordinal regression model, we will combine these $N \times K$ values into a single matrix $\mathbf{M}$, and use $\mathbf{M}_{\cdot k}$ and $\mathbf{M}_{n\cdot}$ to denote $N \times 1$ column and $1 \times K$ row vectors respectively. Each of the L experts then creates their own noisy realisation of this matrix, $\mathbf{Y}^l$ by sampling each row, $\mathbf{Y}^l_{n\cdot}$, from $\mathcal{N}(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l^{-1}\mathbf{I}_K)$. As for $\mathbf{M}$, $\mathbf{Y}^l_{n\cdot}$ represents a $1 \times K$ vector and $\mathbf{Y}^l_{\cdot k}$ an $N \times 1$ vector. Each expert finally assigns the class t_{nl} to the value k that satisfies $y^l_{nk} > y^l_{ni}\ \forall i \neq k$. Denoting the full $N \times L$ matrix of labels as $\mathbf{T}$, this generative procedure corresponds to the following likelihood for the nth datapoint over all L experts

$$p(\mathbf{T}_{n\cdot}, \mathbf{Y}^1_{n\cdot}, \dots, \mathbf{Y}^L_{n\cdot}|\mathbf{M}_{n\cdot}, \boldsymbol{\alpha}) = \prod_{k=1}^{K}\prod_{l=1}^{L}[\mathbf{1}(y^l_{nk} > y^l_{ni}\ \ \forall i \neq k)\mathcal{N}(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l^{-1}\mathbf{I}_K)]^{\mathbf{1}(t_{nl}=k)} \qquad (13)$$

From this likelihood and assuming independent GP priors on each of the K columns of $\mathbf{M}$ and a $\mathcal{G}(v, \lambda)$ prior on each α_l, we can obtain conditional distributions from which we can build a Gibbs sampler with

the following conditional distributions

$$p(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, t_{nl}=k) \propto \mathcal{N}(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l^{-1}\mathbf{I}_K)\mathbf{1}(y^l_{nk} > y^l_{ni}\forall i \neq k) \tag{14}$$

$$p(\mathbf{M}_{\cdot k}|\mathbf{Y}^1_{\cdot k}, \ldots, \mathbf{Y}^L_{\cdot k}, \boldsymbol{\alpha}, \mathbf{C}) = \mathcal{N}\left(\mathbf{M}_{\cdot k}|\boldsymbol{\Sigma}\sum_{l=1}^{L}\alpha_l\mathbf{Y}^l_{\cdot k}, \boldsymbol{\Sigma}\right), \quad \boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \mathbf{I}_N\sum_{l=1}^{L}\alpha_l\right)^{-1} \tag{15}$$

$$p(\alpha_l|\mathbf{M}, \mathbf{Y}^l, v, \lambda) = \mathcal{G}\left(v + \frac{KN}{2}, \lambda + \frac{1}{2}\sum_{k=1}^{K}(\mathbf{Y}^l_{\cdot k} - \mathbf{M}_{\cdot k})^T(\mathbf{Y}^l_{\cdot k} - \mathbf{M}_{\cdot k})\right) \tag{16}$$

As in the ordinal regression case, we can use these distributions to generate samples from the posterior distribution over $\mathbf{M}$, from which it would be possible to make predictions of the label of a new datapoint. If we are not interested in the individual expert precisions, α_l, we can make the assumption $\alpha_l = 1 \ \forall l$ which leads to a simplified representation significantly reducing the number of separate $\mathbf{Y}$ matrices required from L to $\min(L, K)$. Details are provided in appendix B.

More detailed derivations of the conditional distributions and a description of the sampling scheme can be found in appendix A.

2.3.1 Making Predictions

Our principal aim in building a classifier is to be able to make predictions regarding previously unseen data, i.e. we would like to compute

$$p(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{T}, \gamma, v, \lambda) \tag{17}$$

by marginalising over all latent variables in the model. This leads to the following predictive distribution (details given in appendix C)

$$p(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{T}, \gamma, v, \lambda) = \frac{1}{N_{samps}}\sum_{s=1}^{N_{samps}} \mathrm{E}_{p(u)}\left[\prod_{j\neq k}\Phi\left(u + \left(\sum_l \alpha_l^s\right)^{1/2}(m^s_{new,k} - m^s_{new,j})\right)\right] \tag{18}$$

where the s superscript above a parameter denotes the sth sample of that parameter from the Gibbs sampler. The expectation is over the random variable $u \sim \mathcal{N}(0, 1)$ and as such can be easily computed by sampling. This expression is derived in the same manner as the predictive distributions in [14, 10] to which the reader is directed for more information.

2.3.2 Identifiability

As with the ordinal regression model, we must be careful to ensure identifiability of the parameters. Strong identifiability can be ensured by fixing the variance for one of the raters (say α_1). This defines the scale for all classes for rater 1 and hence the scale of $\mathbf{Y}^l$. In turn, this fixes the scale of $\mathbf{M}$ for all raters. Alternatively we can assign proper priors to these parameters that are sufficiently strong to ensure weak identifiability. Whilst in our synthetic experiments we found that weak identifiability was sufficient, we did notice a small but significant increase in convergence time when using real data. Performance of the two approaches will be dependent on the dataset being investigated so it is advisable to try and compare both methods to ensure that the sampler is converging in a satisfactory manner.

2.4 Model scaling

As already mentioned, the model scales linearly with the number experts. Additionally, scaling with the number of classes is linear. For many other multiclass GP approaches (e.g. [19]) it is cubic due to

the inversion of a $KN \times KN$ covariance matrix. Here, we need only invert at worst K separate $N \times N$ covariance matrix is required, and typically only one if class conditional matrices are not required.

Worst case memory and time scalings are N^2 and N^3 respectively. Much effort has recently gone into developing sparse GP approximations (an example for the multinomial probit GP is given in [9]) and there is no reason why such approximations could not be used here if necessary.

2.5 Empty ratings

In many applications generating ratings is expensive. Hence, not all raters will rate every data instance. There may also be too many for any one expert to rate. This simply requires some additional book-keeping. For example, defining $t_{nl} = 0$ if expert l has provided no rating for instance n, in the ordinal regression case the Gibbs sampling distributions in equation 10 must be modified as follows. Firstly, values for y_{nl} will only exist for objects (n) that have been rated by the lth expert. The distribution for $\mathbf{m}$ will be further decomposed thus

$$p(\mathbf{m}|\mathbf{Y}, \boldsymbol{\alpha}) \propto p(\mathbf{m}) \prod_{l=1}^{L} \prod_{n=1}^{N} [\mathcal{N}(y_{nl}|m_n, \alpha_l^{-1})]^{\mathbf{1}(t_{nl}>0)}. \tag{19}$$

Similarly, in the update for α_l, the second term will be decomposed to a summation over the instances rated by expert l. The problem is solved in exactly the same way for classification.

3 Illustrative examples

3.1 Ordinal regression

We first use synthetic data to investigate the convergence properties of strongly and weakly identifiable Gaussian process models. Following [6] we generated $N = 200$ datapoints x_n from $\mathcal{N}(x_n|0, 1)$. True latent trait values for each of $L = 4$ raters were then generated according to $m_{nl} = 1 - 2x_n + \epsilon_{nl}$ where $\epsilon_{nl} \sim \mathcal{N}(0, \alpha_l^{-1})$ and the precision values for the four raters were $\alpha_l = [1, 0.5, 2, 4]$. A radial basis covariance function was used, defined as

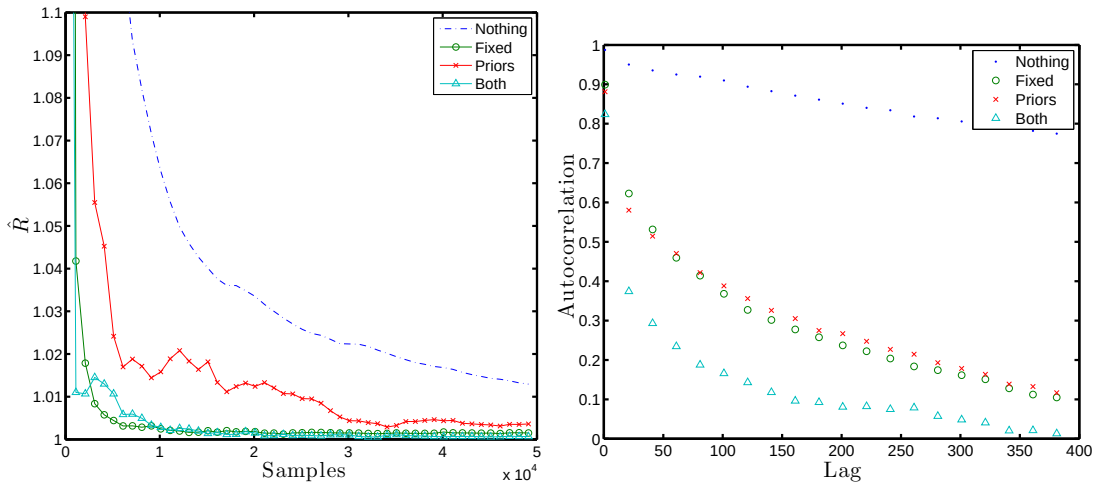
$$C(x_i, x_j) = \exp(-\gamma(x_i - x_j)^2) \tag{20}$$

with $\gamma = 1$. With $K = 5$, the true thresholds were sampled from $\mathcal{N}(0, 1)$ subject to the necessary order constraints. To assess the performance of weak and strong identifiability we ran the sampler in four configurations. Firstly, an unidentifiable model with no fixed parameters and diffuse priors (a uniform prior over the entire real line for b (subject to the necessary order constraints $b_{10} < b_{11} < \cdots < b_{1K}$) and a Gamma prior with parameters $a = b = 0$ for the precisions ($\mathcal{G}(\alpha_l|0, 0)$)). Secondly, a strongly identifiable model with $\alpha_1 = 1$ and $b_{11} = 0$ fixed. Thirdly, a weakly identifiable model with no parameters fixed and proper priors on these parameters ($b \sim \mathcal{N}(0, 1)$, again subject to the required ordering and $\alpha \sim \mathcal{G}(1, 1)$). Finally, we combined the two methods by fixing some values and placing informative priors on the non-fixed values. Figure 3(a) gives the $\hat{R}$ statistic that compares inter- and intra chain variances across multiple chains ([8]) for the four configurations. We notice that convergence is fastest in the two cases where parameters are fixed (strong identifiability, labeled 'fixed' and 'both'), slower in the weak case ('priors') and very slow in the un-identifiable case as the parameters are just wandering in a coupled manner through the space (see figure 4(b)). In figure 3(b) we show the autocorrelation for α_2. Strong or weak identifiability alone ('fixed' and 'priors') offer similar performance and large improvements are seen when they are combined. The autocorrelation is still rather high, suggesting thinning of the output samples is required - a characteristic of models of this type - [6] investigates methods for speeding up

sampling for a linear model and investigating similar techniques for the GP prior is an interesting avenue for future investigation.

Given that fixing one threshold and one precision causes only a small loss in interpretability, it would seem that in this case combining strong and weak identifiability is a sensible choice. However, given the small time required to generate a large number of samples (less than 10 minutes for 50000 on a standard desktop machine in this example), if one is concerned with the reduction in interpretability, the weakly identifiable model is still feasible. Finally, in figure 4, we see trace plots for the precisions in the un-identifiable (a) and combined strongly and weakly identifiable (b) cases.

Figure 3 about here



(a) (b)

Figure 3: Convergence statistics for the toy example. (a) the $\hat{R}$ statistic of [8] that compares inter- and intra- chain variances. (b) the autocorrelation.

Figure 4 about here

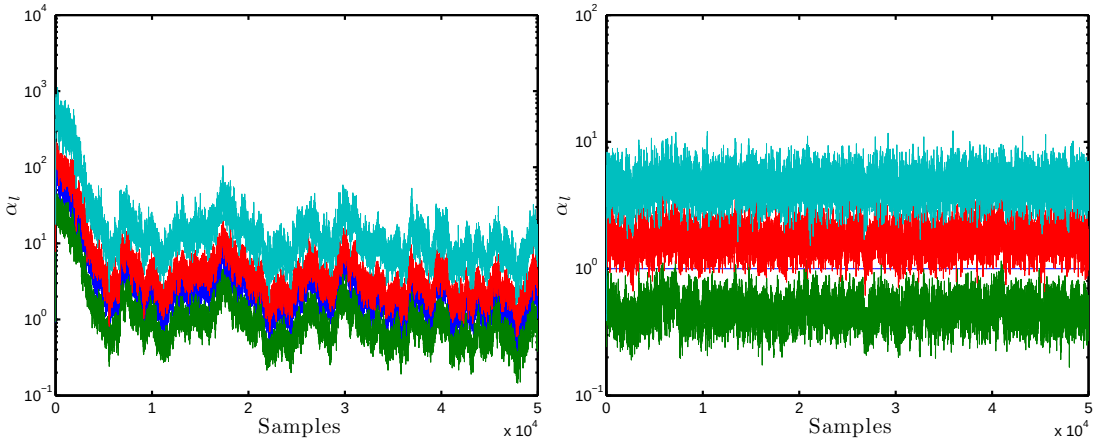


Figure 4: Trace plots for the synthetic example in the unidentifiable (a) and when using informative priors and fixing values (b).

We now provide a comparison with the linear regression model of [13] and a model using no covariate information. The dataset is generated as follows. $N = 50$ data points are sampled with dimension M from $\mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_M)$ and a *true* linear weight vector $\boldsymbol{\beta}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_M)$. For each of $L = 5$ experts, we generate $\mathbf{Y}_{\cdot l}$ from a Gaussian with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance $\alpha_l^{-1}\mathbf{I}_N$ (where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$). For each rater, we sample thresholds from $\mathcal{N}(b_{lk}|0, \sigma_b^2 = 2)$ (subject to the necessary order constraints) and hence generate labels. We repeat this experiment for $M = 20$ and $M = 50$, and three different priors from which the α_l are drawn, corresponding to decreasing expected noise level. We will compare five algorithms (listed in the order they appear in the figure) - GP multi-rater algorithm with no covariate information ($\mathbf{C} = \mathbf{I}_N$) (red), the algorithm of [13] (with the reversed conditional distributions to ensure identifiability) (black), linear regression model with informative prior on $\boldsymbol{\beta}$ (blue) (this effectively corresponds to the distribution from which the data were drawn), GP multi-rater algorithm with all precisions fixed to 1 (green) and the full GP multi-rater algorithm (magenta). In all cases (except the algorithm of [13] where the code was used exactly as provided) we used the combination of strong and weak identifiability that performed well in the previous example. The models used are summarised in table 1.

Name	Description	Color
GP ($\mathbf{C} = \mathbf{I}$)	GP multirater algorithm with no covariate information	Red
Johnson	The algorithm of [13] with reverse conditionals to ensure identifiability	Black
Linear	Linear regression with informative prior on β	Blue
GP (fixed)	GP multirater algorithm with precisions fixed to 1	Green
GP	Full GP multirater algorithm	Magenta

Table 1: Models compared in linear ordinal regression example (shown in order of figure 5).

The results can be seen in figure 5 where the y-axis corresponds to the Spearman's rank coefficient between the posterior mean latent function value and the true rankings, computed from $\mathbf{X}\boldsymbol{\beta}$. The higher the coefficient, the better the model performance with a ranking of $\rho = 1$ corresponding to an exact match. 20 datasets were created with each combination of covariate set size and noise prior. For each

dataset, each algorithm was run 10 times to help assess chain convergence.

Figure 5 about here

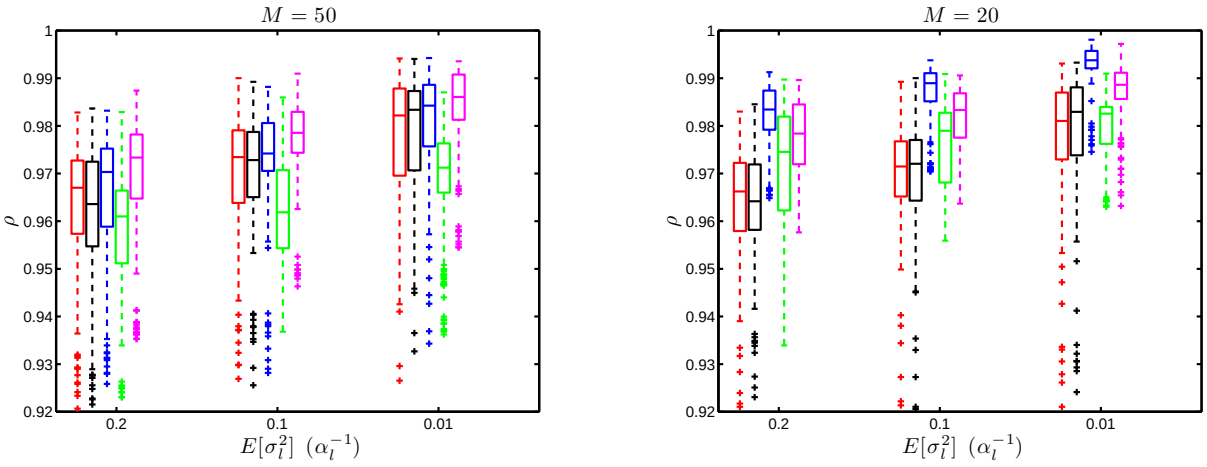


Figure 5: Results from linear ordinal regression example. The boxes denote (from left to right), GP multirater with no covariate information (red), algorithm of Johnson et al. with reverse conditionals (black), linear regression model with informative prior on $\boldsymbol{\beta}$ (blue), GP multirater algorithm with precisions fixed to 1 (green), full GP multirater algorithm (magenta).

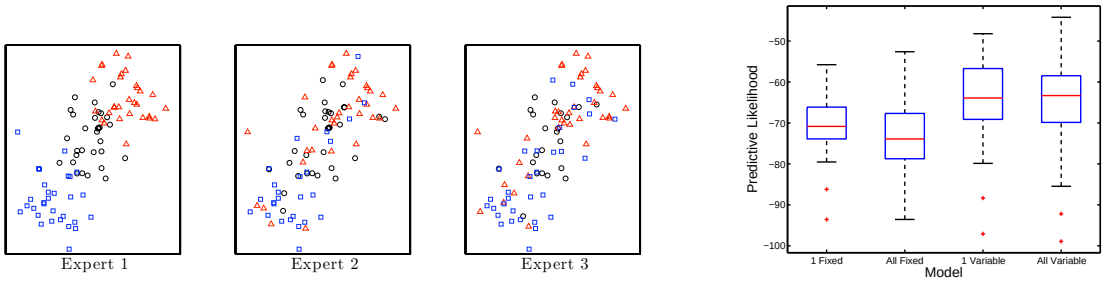
The first thing that is obvious from figure 5 is that the performance improves as noise is decreased (left to right on the x-axis). Also, when the size of the feature set equals the number of training examples ($M = N = 50$), the full GP (magenta) outperforms all other methods including the linear model. This shows the benefit of the GP method when faced with problems with large numbers of covariates. For the small covariate set, the linear model with the proper prior outperforms all other algorithms. This is to be expected as it is exactly the model that created the data. There are other interesting comparisons that can be made. Firstly, the full GP model always outperforms the GP model with precisions fixed at 1 (green). Secondly, both the full GP model and the linear model with a proper prior always outperform the baseline model that does not include covariate information (red). However, the algorithm of [13] does not, and is occasionally worse than the baseline. This highlights the importance of the prior on the latent function - in this example, placing a proper prior on the regression coefficients and fixing two of the parameters appears to be far more effective than the alternative. Whether or not prior information regarding the regression coefficients is available, the GP prior looks like a promising alternative. It is worth noting that this experiment is biased rather heavily towards the linear model with proper prior as the true prior distribution from which the regression coefficients were sampled is known. Hence, in reality, the performance of the linear model may not be as good.

3.2 Classification

As with ordinal regression, we will first demonstrate the performance on a toy problem. We generate a toy dataset by sampling 30 datapoints from each of three classes defined as Gaussians with means $[0, 0], [2, 2], [-2, -2]$ and unit diagonal covariance matrices. We provide the first expert with the true labels and then add two additional experts with noisy labels, where each label is either equal to the true

label with probability 0.5 or otherwise drawn randomly. The performance of a standard classifier with only the labels from the first expert provides the best performance possible. Our aim in this example is to see how much performance drops when the labels from the other experts are added. Ideally, if the model performs well, the decrease in performance will be insignificant. We will try two variations of our multi-expert model. In the first, we will fix the expert-specific precision values to $\alpha_l = 1$, whilst in the second model we will allow the precisions to vary, with the prior distribution set as a Gamma distribution with parameters $v = 1, \lambda = 1$. It is unreasonable to hope that the performance of either of these models will equal that of the classifier trained on the true labels. However, we would expect the model with fixed precisions to perform worse than the model where the precisions can vary as the latter model is able to assign less certainty to the noisy models. A Gaussian covariance function is used ($C(x_i, x_j) = \exp(-\gamma(x_i - x_j)^2)$) with $\gamma = 1$.

Figure 6 (a) and (b) about here



(a) (b)

Figure 6: (a) Toy classification data labeled by three experts. (b) Comparison of predictive log likelihoods for the toy classification examples. '1 Fixed' is a classifier trained on the true labels from expert 1 with α fixed at 1. 'All Fixed' also includes labels from experts 2 and 3 and again α is fixed. '1 Variable' is trained with just labels from expert 1 but now α is inferred and 'All Variable' is trained on all of the labels with α inferred.

We can see the data and the performance of the three models in figure 6 where the boxplots show the predictive likelihoods over 50 randomly sampled datasets. '1 Fixed' corresponds to a standard classifier trained on just the labels from expert 1 (the truth) with $\alpha = 1$ fixed. 'All Fixed' is a classifier trained on all of the labels with $\alpha_l = 1 \ \forall l = 1 \dots L$. '1 Variable' is trained on just the labels from expert 1 but with α inferred whilst 'All Variable' is trained on all labels with α_l inferred. We see that in both cases (with α fixed or inferred), the inclusion of the noisy labels leads to a drop in the median predictive likelihood (-70 to -73 when α is fixed, -64 to -65 when inferred), however, neither drop is very large. This suggests that the model handles the noisy labels reasonably well. The drop is particularly small when α is inferred as the model is able to assign less certainty to the unreliable labelers. We can assess the significance or otherwise of the changes using a paired t-test. Comparing the fixed and variable models we obtain $p = 1.4 \times 10^{-4}$ in the single expert case and $p = 6.1 \times 10^{-5}$ when all labels are used suggesting that there is indeed an improvement when we allow the model to infer precisions. Comparing using the first set of labels versus all of the labels in the two cases gives $p = 0.0612$ in the fixed case and $p = 0.5462$ in the inferred case. Whilst one cannot reject the null hypothesis at 5% in either case, the decrease in performance when noisy labels is added seems to be much smaller when the model is free to infer precisions, as one would expect. In one arbitrary run from this example, the posterior mean values for the three precisions were $\boldsymbol{\alpha} = [3.85, 0.14, 0.26]$. It is clear that much more certainty has been assigned

to labeler 1.

Finally we found that the priors used were strong enough to ensure posterior identifiability - fixing one of the precision parameters in the 'All variable' model resulting in strong identifiability made no observable difference to chain convergence.

4 Classification example - labeling clinical reports

We now turn to a real dataset from the 2007 Computational Medicine Center (CMC) Medical NLP Challenge[4]. The original data consisted of anonomised medical records from a children's hospital consisting of two parts, the medical impression and medical history. The medical history briefly describes patients' prior complaints, while the impression describes the results of the current examination. Each record also includes ICD9 codes[5] assigned by three different companies. We have removed all classes for which there is only one instance, leaving a total of 28 classes. In some instances, individual companies suggest more than one label. Extending the current method to be able to handle multi-label data such as this is an interesting avenue for future work. However, at the moment, we discard these instances, leaving a total of 497 reports.

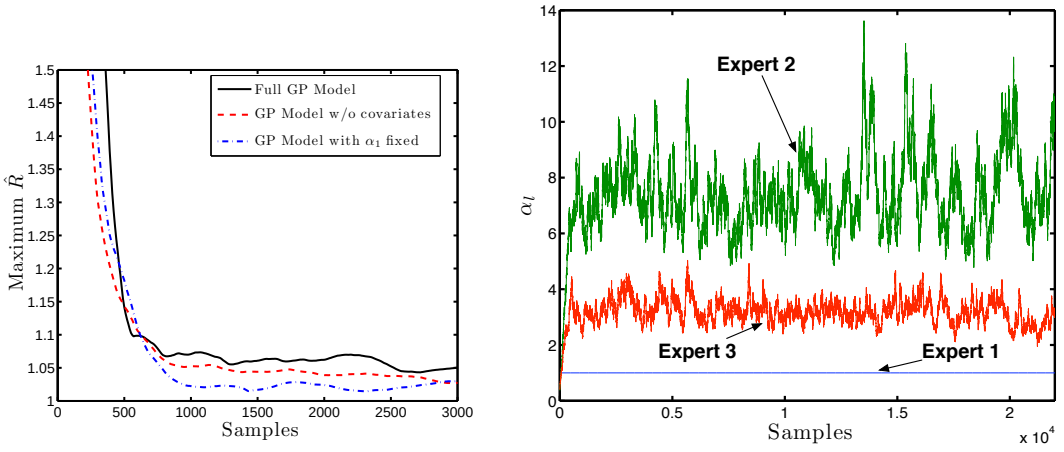
Despite the existence of guidelines as to how the data should be labeled, we see a very high level of disagreement between the three experts. Figure 8(a) shows the number of examples assigned to twenty of the largest classes. Notable classes are 13, where expert 1 makes 30 assignments whilst expert 3 makes none and class 3 where experts 2 and 3 make many more assignments than expert 1. This demonstrates the difficulty of the labeling task due to the subjective nature of the data and motivates the development of methods capable of dealing with such diversity.

The clinical impression and history were each represented as a document-feature co-occurrence matrix. The features were all of the words which occur in the records excluding a standard list of English stop words[6]. The words were also normalised by converting to lower case. In addition, the patient age descriptions were standardised by converting all numbers to digits and concatenating the description words into a single token (e.g. `Four month old` would become `4_month_old`). This left about 1500 unique tokens, a far greater number than the number of data instances. In our experiments we used the history portion of the data as this was the most informative. Incorporating the impression information too is an interesting avenue for future work, possibly through a convex combination of GP covariance functions as described in [10].

In this experiment, a linear covariance function ($C(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i^T \mathbf{x}_j$) was used after the input vectors were normalised to have unit norm. The Gamma prior had parameters $v = 1, \lambda = 1$. To monitor convergence we compared the weakly identifiable GP model with a strongly identifiable model and with a weakly identifiable model without any covariate information (the covariance matrix, C is equal to an identity matrix). Figure 7 shows the evolution of the $\hat{R}$ statistic for the precision parameters (the maximum value is taken across the 3 (2 in the strongly identifiable case) parameters at each sample). We see that there is a small but significant difference in convergence between the strongly and weakly identifiable cases. Also in figure 7 we can see a trace plot of the parameters in the strongly identifiable models. We can see that there is a rather high level of autocorrelation and hence thinning is required. Bearing this in mind, in our experiments we use 2000 burn in samples and then a further 20000 samples

[4]Data and detailed description on the CMC website http://www.computationalmedicine.org/challenge/index.php
[5]http://icd9cm.chrisendres.com/index.php
[6]ftp://ftp.cs.cornell.edu/pub/smart/english.stop

from the posterior, thinned by a factor of 200 to give 100 final samples from which to compute predictive distributions.



(a) (b)

Figure 7 about here

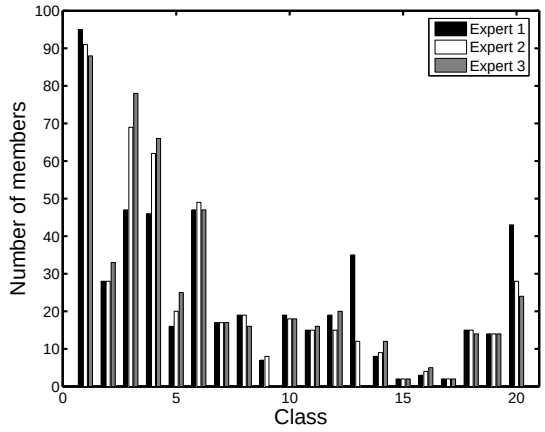
Figure 7: (a) Monitoring convergence of the Gibbs sampler for the precision parameters for the weakly identifiable full GP model, a weakly identifiable GP model without covariate information and the strongly identifiable model obtained by fixing α_1. (b) Trace plot of the precision parameters for the strongly identifiable GP model.

Evaluation of the methods is difficult as we do not have a ground truth with which to compare. In [15], a modified receiver operating characteristic (ROC) curve is used. This is possible due to the ordinal nature of their four categories. This is not possible in our case as the classes have no ordering and true multi-class ROC analysis is somewhat unwieldy. Therefore, we will consider two more simple measures here. The first assumes that we would like the outcome of our algorithm to be single predictions of the class of a new instance and so we will compare the class predicted with highest probability with the majority given by the labelers (in only one instance is there not a majority). This is a somewhat crude measure - if we are going to test on the majority, why not simply train on the majority? However, our results below show that we do see an improvement in the prediction on the majority when training with all labels. Secondly, we assume that rather than a single prediction, we would like our algorithm to provide a predictive probability distribution over all classes that reflects the true certainty/uncertainty of the problem. Such a distribution might be a useful low-dimensional representation of the report in a database against which to conduct searches. Specifically, we would like all examples on which the labelers agree to be predicted correctly with very high probability. Test points for which there is disagreement should be predicted with less certainty and additional predictive weight given to any minority class.

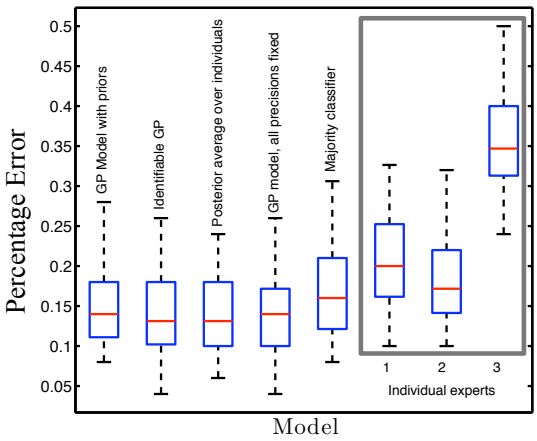
Considering the first of these measures, results from 10-fold cross validation, repeated with 10 different partitionings of the data can be seen in figure 8(b). From left to right, the different models correspond to: The weakly identifiable GP, the strongly identifiable GP, a *post-hoc* posterior averaging procedure over classifiers built from the individual experts' labels, the GP model with all precisions fixed to unity, a classifier trained on the majority label and classifiers trained on the labels from the three individual experts. We see that the best performance is shared between the multiple-expert algorithm (with fixed or inferred precisions) and predictions based on the posterior average of the three individual experts.

The performance difference between these two methods is not statistically significant (paired t-test, 5% significance value) although each of them is significantly better than the remainder. When training is performed with the majority labels, performance is indistinguishable from that of classifiers trained with just the labels from expert 2. Expert 3 on the other hand is performing rather badly. Interestingly, expert 2 agrees with expert 3 more often than it does with expert 1 so it cannot be the case that this performance difference is just due to experts 1 and 2 always making the majority. The regular agreement between experts 2 and 3 can also be observed in the posteriors for the precision parameters, α, figure 8(c). Here we see that the precision for expert 2 is much higher than those for experts 1 and 3. This is not surprising as we have already observed that a classifier trained with the labels from expert 2 alone performs well with respect to the majority labeling. However, comparing the performance of classifiers trained on labels from experts 1 or 3, we might predict that the precision for expert 1 in the combined classifier would be higher than that for expert 3. In fact, the opposite is the case. The fact that all three methods that use all labels (unidentifiable and identifiable multi-expert GPs and posterior averaging) perform significantly better at predicting the majority label than the classifier trained on the majority label appears somewhat counterintuitive. However, we suspect that this is due to the increased information available to these classifiers - it seems reasonable to assume that data instances for which the experts disagree should provide some information about each of the classes mentioned - if we use only the majority label, we are reducing the set of training instances associated with each class.

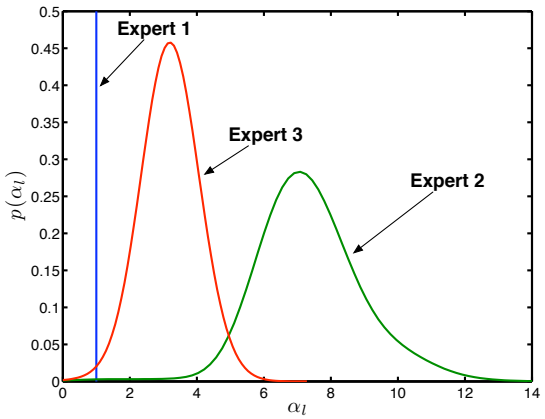
Given the results in our synthetic example, the fact that the GP algorithm with inferred precisions does not outperform that with fixed precisions seems surprising. This may well be an artifact of this dataset and comparisons on other datasets, possibly with more experts is an avenue for future work.



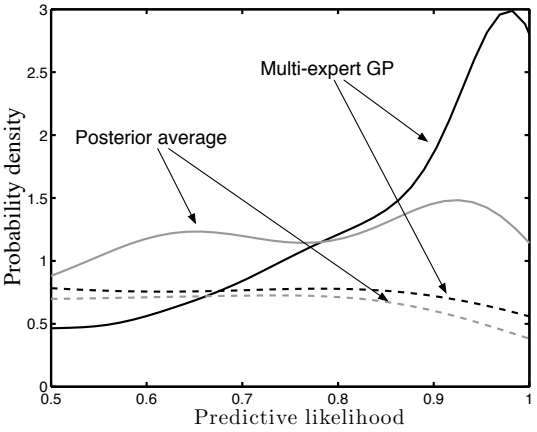
(a) Distribution of assignments by the three experts over 20 of the largest classes in the clinical competition data.



(b) Majority test error



(c) Precision posteriors (note that the precision for expert 1 is fixed at 1).



(d) Empirical distribution of predictive probabilities in test cases. x-axis corresponds to the probability assigned to the true class, y-axis the empirical density of this probability. Solid lines correspond to majority labels, dashed to minority (where they existed).

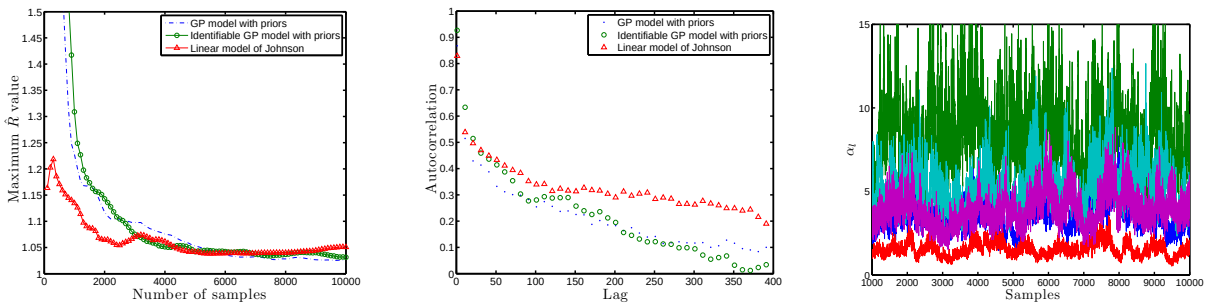
Figure 8 (a,b,c,d) about here

Figure 8: (a)Distribution of assignments by the three experts over 20 of the largest classes in the clinical competition data, (b) accuracy (0-1 loss) with respect to majority labels, (c) example posterior distribution over the precision for the three experts in the strongly identifiable case and (d) comparison of predictive likelihoods in the strongly identifiable case (black lines) with a model built from averaging the predictions over the three experts (grey lines), solid curves correspond to majority labels, dashed to minority (where they existed).

Our second method of evaluation (quality of predictive distributions) is more difficult but we can gain some insight by looking at predictive probabilities for the cases when all labelers agree, and cases when they don't. Figure 8(d) shows a comparison of the distributions of predictive likelihoods between the multi-expert and the posterior average of the individual classifiers for test points in which the labelers

agree (solid curves) and disagree (dashed curves). We can see that the multi-expert model is able to predict both majority and minority labels with more certainty than the posterior average of the individual classifiers as it has more probability mass towards the higher end of the x-axis. There appears to be a smoothing effect inherent in the posterior averaging - the combination treats predictions from expert 3 with the same weight as those from experts 1 and 2 and it seems likely that this flattening effect will get worse as the number of experts increases. This evidence suggests that the multi-expert algorithm provides a more natural and reliable framework for encapsulating the uncertainty inherent in this data. It should be noted that the posterior averaging performance could probably be improved with more sophisticated combination schemes. However, the multi-expert algorithm seems a more sensible option as it is able to implicitly optimise this combination through the expert-specific precision parameters.

5 Ordinal regression application - essay grading



(a) (b) (c)

Figure 9 (a),(b) and (c) about here

Figure 9: Monitoring convergence of the precision parameters in the ordinal regression example using (a) the $\hat{R}$ statistic and (b) the autocorrlation. (c) trace plot of the precision parameters for the full GP model.

Our final example uses the dataset discussed in [13]. The data consists of features derived from 198 essays (number of spelling mistakes, average word length etc) and ratings for each essay (between 1 and 10) from each of five experts. When analysing this data there are three obvious aims. Firstly, we may wish to generate a relative ranking of all of the papers, incorporating the opinions from all of the experts. Secondly, we may wish to quantify the variance of each of the raters - it is clearly useful to know how consistent any particular rater is. Finally, we may wish to produce an automated grader. We will consider the first two problems here. For discussion relating to the third, the reader is directed to [13] and [12].

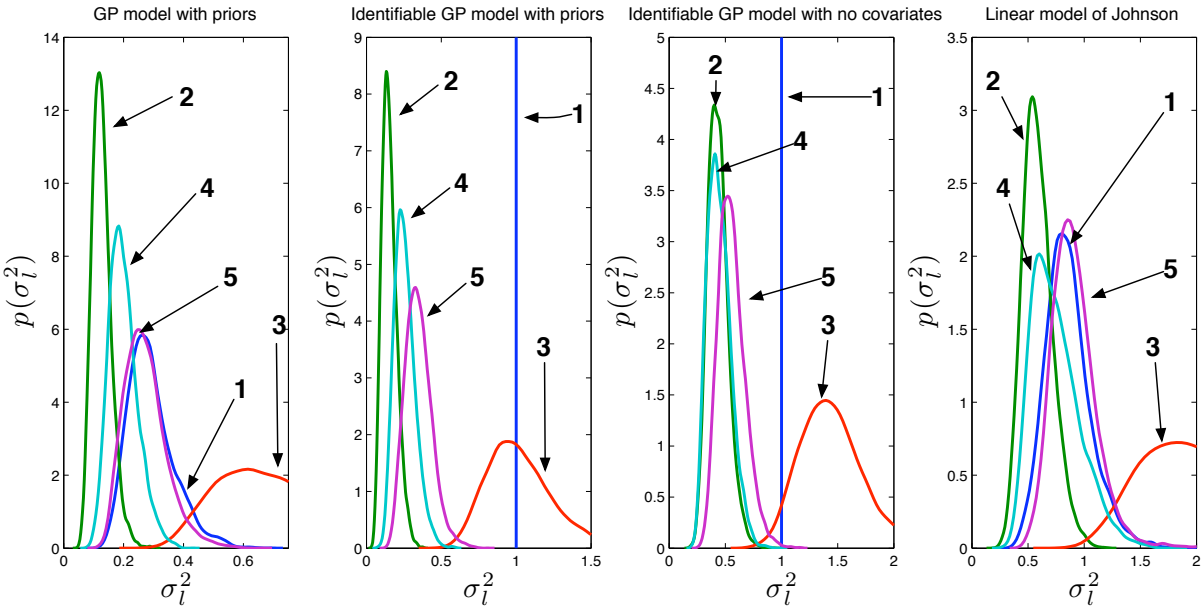


Figure 10 about here

Figure 10: Left to right: Variances obtained for the five examiners in the essay dataset for the weakly identifiable GP model ($1/\alpha_l$), the strongly identifiable GP model, a strongly identifiable GP model with no covariate information ($\mathbf{C} = \mathbf{I}$) and the regression model of [13]. The same effective prior on variance was used in all cases. The ordering of experts from the most (expert 2) to the least (expert 3) consistent is conserved except in the final case where fixing the values of parameters changes the posterior distribution from which we're sampling.

In the following experiment, we use a linear covariance function ($C(x_i, x_j) = x_i^T x_j$) where each input vector is normalised to have unit norm. As in the classification example, we assess convergence and check for possible problems with identifiability using the $\hat{R}$ statistic (see e.g. [8]) for the precision parameters (α_l). Figure 9 shows the evolution of the maximum value (over the five parameters) of this statistic over ten independent chains, the autocorrelation of α_2 and a trace plot of the precision parameters from the full, weakly identifiable GP model. As well as convergence statistics for the weakly identifiable GP model, we also show the evolution for a strongly identifiable GP model with with priors and the linear regression model of Johnson ([12, 13]). We see that the two GP models are roughly consistent and appear to have converged to a lower $\hat{R}$ value after 10000 samples than the linear model. Similarly, in figure 9(b) the two GP models exhibit similar autocorrelation behaviour, in both cases better than that for the linear model. In the following experiments, we generate 10000 samples from our Gibbs sampler, discarding the first 5000 as a burn-in. For comparison, we used the code provided by the authors of [12] exactly as provided, initialised with the results of their covariate free model. The parameters of the prior on the expert precisions is Gamma with parameters $v = 1, \lambda = 1$ and $\sigma_b^2 = 1$.

In figure 10, we show the marginal posterior densities for the variances (we have plotted variances rather than precisions for ease of comparison with the regression model) of each rater for the three models described in the previous paragraph and a GP model with no covariate information (strongly identifiable with informative priors). The effective prior on variance was the same for all models and so it is interesting to note that, whilst the ordering from least to most consistent raters is roughly similar,

the expert variances for the GP models are consistently smaller than those for the regression model - most likely due to the different priors over the latent trait function. The posterior densities for the three GP models are very similar, suggesting that the posterior is indeed identifiable despite the fact that the likelihood is not. In the strongly identifiable case, we see that fixing the precision for rater 1 causes the ordering to change slightly (it now has similar precision to rater 3 whereas before it was higher (variance lower)). This is not surprising - placing a restriction on the thresholds will generally mean that the first term in the likelihood

$$p(t_{nl}, y_{nl}|m_n, \alpha_l) = \mathbf{1}(b_{l,t_{nl}-1} < y_{nl} \leq b_{l,t_{nl}})\mathcal{N}(y_{nl}|m_n, \alpha_l^{-1}) \tag{21}$$

is constant and hence the second term will need to be more flexible. If the values of b_{11} and α_1 are fixed to the posterior mean from the weakly identifiable model, the original ordering is recovered (result not shown).

Whilst both methods seem to perform similarly here, it is important to remember that with the GP prior, we have not had to fit specific regression parameters for each feature. This is beneficial in problems with large numbers of covariates relative to the number of examples. In such a regime, a linear model would require regularization to avoid overfitting. There is no such issue with the GP prior. For example, we could use the essays themselves in this example rather than features derived from them. We do however lose any information regarding how important each covariate is although relevances for each could be obtained using an automatic relevance detection (ARD) covariance function as described in [9].

Figure 11 about here

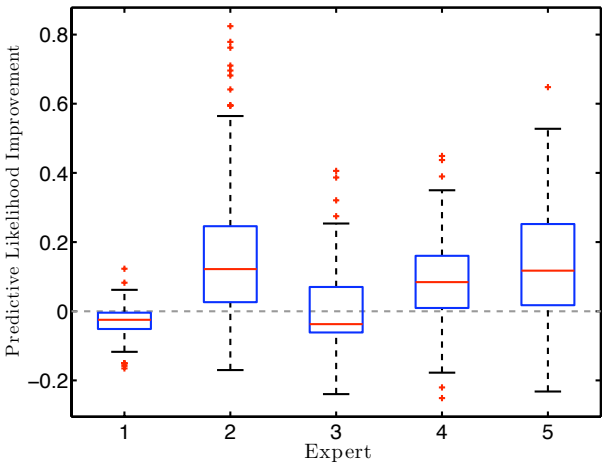


Figure 11: Gain in predictive likelihood for predictions by the 5 experts when the multi-rater model is used rather than individual ordinal regression models.

Finally, we can use the predictive capability of the model to determine to what extent analysing the data in a joint manner is better than simply fitting individual ordinal models. Using a 10-fold cross validation procedure, we can compute predictive likelihoods for each expert over the held-out data within the multi-rater model and in individual ordinal regression models. All settings were the same as in the previous experiments with this data above. In figure 11 we have plotted the improvement in predictive likelihood for the multi-rater model over individual models for the 5 experts. We can see that the multi-rater model is able to predict the behavior of individual raters better than models just based on their

individual responses. For the other 2, the decrease in performance is rather low. This is an interesting and promising conclusion - it appears that the general consensus between the raters enables the model to make correct predictions with greater confidence. It is not surprising to notice that the two experts for which predictive performance is worst are those with the lowest associated precision (highest variance) values (c.f. figure 10).

6 Conclusions

In this paper we have introduced semi-parametric models able to handle both classification and ordinal regression data that has been subjectively labeled by a number of different experts. The classifier is built upon the multinomial probit classifier of [9] and as such remains practical when the number of classes and (with the use of sparse approximations) the number of training examples are high. Additionally, it can be extended to semi-supervised problems ([14]) and the combination of different data sources through composite covariance functions ([10]). To the best of our knowledge, this is the first non-parametric classifier that can accommodate data instances labeled subjectively by several experts.

Unfortunately, as the latent trait functions for both ordinal regression and classification have no implicit scale, the likelihood functions for both models are unidentifiable. We have investigated two options for overcoming this problem. Firstly, ensuring strong identifiability by fixing particular parameter values (a subset of thresholds for ordinal regression and one precision parameter for classification) and secondly, following [11], ensuring weak identifiability by placing suitable priors on all parameters. In our experiments, we did not find one strategy which consistently outperformed the others. Strong identifiability seems to be the safest route, however there is some loss in interpretability and given the relatively low computational cost of generating samples, the potential cost of generating more samples to overcome the higher autocorrelation in the weakly identifiable case (particularly for classification) may not be too much of a burden. In practice, a combination of the two strategies may be the best approach. The inclusion of covariate information can lead to difficulties in ensuring that a proper prior is placed on the latent functions ([13]), particularly if little prior knowledge is available regarding the problem domain. This is elegantly overcome through the use of a Gaussian process prior on the latent function. The prior is proper and does not require any parametric assumptions. In our investigations we have assumed Gaussian noise. In many applications, this may not be suitable and following [2] in extending the model to to the t-link rather than the probit link may be an interesting avenue for future investigation. Similarly, it may be possible to improve the efficiency of the sampler (particularly for the cut-points) in the manner described in [1].

Presently, inference is restricted to using Gibbs sampling to sample from the posterior over model parameters. This is computationally rather intensive and approximations may be useful in large problem domains. Under certain restrictions ($\alpha_l = 1 \; \forall \; l$) it is possible to create a variational (mean field) approximation to the classification model along the lines of that shown in [9], although in the more general case and for ordinal regression, development of suitable approximations is an interesting area for future work. Expectation propagation (as used in [4]), may be a practical alternative to a variational approximation in the ordinal regression case.

We anticipate the model being of particular use in the text domain where the GP prior is not restricted by the very large number of features and the labeling of training instances is often rather subjective. Explicitly modeling this uncertainty in the classification algorithm is likely to be more efficient and possibly more effective than forcing the labelers to come to an agreement in contentious cases. We have shown that our model outperforms a similar GP trained on the majority labeling for a subset of the data from a recent classification competition. Raw classification performance is comparable with taking a posterior

average over classifiers trained on the individual labels but predictive likelihoods are considerably higher. We anticipate a similar improvement in performance in other problems where the task of labeling is difficult and open to interpretation. Such datasets are becoming increasingly common, particularly in the biological domain.

In the ordinal regression case, we see results on a benchmark dataset that are in agreement with results produced by a similar, parametric model [12]. This data is of reasonably low dimension and investigation with more complicated data is an avenue for future work. For example, it would in theory be possible to use the actual essays as our data rather than features derived from them.

Matlab code to reproduce the results presented in this paper is available from the authors on request.

7 Acknowledgments

SR is supported by EPSRC grant EP/C010620/1 (Stochastic modelling and statistical inference of gene regulatory pathways: integrating multiple sources of data). MG is an EPSRC advanced research fellow (EP/E052029/1) and this work is also partly supported by the EPRSC through EP/F009429/1 - Classifiers in medicine and biology (Advancing machine learning methodology for new classes of prediction problems). TP is supported by a PhD studentship from Scottish Enterprise.

References

[1] James Albert and Siddhartha Chib. Sequential ordinal modeling with applications to survival data. *Biometrics*, 57:829–836, 2001.

[2] JH Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc*, 88(422):669–679, 1993.

[3] S Bickel, U Brefeld, L Faulstich, and J Hakenberg. A support vector machine classifier for gene name recognition. *BMC Bioinformatics*, Jan 2004.

[4] W Chu and Z Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1–48, 2005.

[5] K Cohen, L Fox, P Ogren, and L Hunter. Corpus design for biomedical natural language processing. *Proceedings of the ACL-ISMB Workshop on Linking Biological ...*, Jan 2005.

[6] M.K. Cowles. Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6:101–111, 1996.

[7] AP Dawid and AM Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[8] A Gelman, J Carlin, H Stern, and D Rubin. Bayesian data analysis. *Chapman&Hall*, Jan 2004.

[9] M Girolami and S Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, Jan 2006.

[10] M Girolami and M Zhong. Data integration for classification problems emplying gaussian process priors. *Advances in Neural Information Processing Systems*, 21, 2007.

[11] V Johnson. An alternative to traditional gpa for evaluating student performance. *Statistical Science*, Jan 1997.

[12] V Johnson and J Albert. Ordinal data modeling. *books.google.com*, Jan 1999.

[13] VE Johnson. On bayesian analysis of multirater ordinal data: An application to automated essay grading. *J Am Stat Assoc*, 91(433):42–51, 1996.

[14] S Rogers and M Girolami. Multi-class semi-supervised learning with the e-truncated multinomial probit gaussian process. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 1:17–32, 2007.

[15] P Smyth, U Fayyad, M Burl, P Perona, and P Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, 7, 1995.

[16] JS Uebersax. Statistical modeling of expert ratings on medical treatment appropriateness. *J Am Stat Assoc*, 88(422):421–427, 1993.

[17] Y Versley. Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-) reference. *Ambiguity in Anaphora Workshop Proceedings*, 2006.

[18] W J Wilbur, A Rzhetsky, and H Shatkay. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356–356, 2006.

[19] CK Williams and D Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

A Gibbs sampler details

A.1 Ordinal regression

The full posterior is given by

$$p(\mathbf{m}, \mathbf{Y}, \boldsymbol{\alpha}, \mathbf{b}|\mathbf{T}, \mathbf{C}, \sigma_b^2, v, \lambda) \propto p(\mathbf{m}|\mathbf{C}) \prod_{l=1}^{L} \left[p(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l) p(\alpha_l|v, \lambda) p(\mathbf{b_l}|\sigma_b^2) \prod_{n=1}^{N} p(t_{nl}|y_{nl}, \mathbf{b}_l) \right]. \tag{22}$$

Taking each component in the posterior in turn, starting with the individual components of $\mathbf{Y}$, y_{nl} we have

$$\begin{aligned} p(y_{nl}|\dots) &\propto p(y_{nl}|m_n, \alpha_l) p(t_{nl}|y_{nl}, \mathbf{b}_l) & (23)\\ &\propto \mathcal{N}(y_{nl}|m_n, \alpha_l^{-1}) \mathbf{1}(b_{l(t_n-1)} < y_{nl} \leq b_{lt_n}). & (24) \end{aligned}$$

For the trait variable, $\mathbf{m}$ we have

$$\begin{aligned} p(\mathbf{m}|\dots) &\propto p(\mathbf{m}|\mathbf{C}) \prod_{l=1}^{L} p(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l) & (25)\\ &\propto \mathcal{N}(\mathbf{m}|\mathbf{0}, \mathbf{C}) \prod_{l=1}^{L} \mathcal{N}(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l^{-1}\mathbf{I}_N) & (26)\\ &= \mathcal{N}(\mathbf{m}|\Sigma \sum_{l=1}^{L} \alpha_l \mathbf{Y}_{\cdot l}),\ \Sigma = \left(\mathbf{C}^{-1} + \mathbf{I}_N \sum_{l=1}^{L} \alpha_l \right)^{-1}. & (27) \end{aligned}$$

For the precisions,

$$\begin{aligned} p(\alpha_l|\dots) &\propto p(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l) p(\alpha_l|v, \lambda) & (28)\\ &\propto \mathcal{N}(\mathbf{Y}_{\cdot l}|\mathbf{m}, \alpha_l^{-1}\mathbf{I}_N) \mathcal{G}(\alpha_l|v, \lambda) & (29)\\ &= \mathcal{G}\left(v + \frac{N}{2}, \lambda + \frac{1}{2}(\mathbf{m} - \mathbf{Y}_{\cdot l})^T (\mathbf{m} - \mathbf{Y}_{\cdot l}) \right). & (30) \end{aligned}$$

Finally, the thresholds

$$\begin{aligned} p(b_{lk}|\dots) &\propto \prod_{n=1}^{N} p(t_{nl}|y_{nl}, \mathbf{b}_l) p(b_{lk}|\sigma_b^2) & (31)\\ &\propto \prod_{n=1}^{N} \mathbf{1}(b_{l,t_n-1} < y_{nl} \leq b_{l,t_n}) \mathcal{N}(b_{lk}|0, \sigma_b^2) & (32)\\ &\propto \mathcal{N}(b_{lk}|0, \sigma_b^2) \mathbf{1}\left(\max_{t_{nl}=k} y_{nl} \leq b_{lk} < \min_{t_{nl}=k+1} y_{nl} \right). & (33) \end{aligned}$$

The Gibbs sampler proceeds as follows

1. Initialise $\mathbf{m} \sim \mathcal{N}(\mathbf{m}|\mathbf{0}, \mathbf{C})$, $\alpha_l \sim \mathcal{G}(\alpha_l|v, \lambda)$ and $b_{lk} \sim \mathcal{N}(b_{lk}|0, \sigma_b^2)$ (subject to the necessary order constraints);
2. Sample a new value for each $\mathbf{Y}_{\cdot l}$ from equation 24;
3. Sample a new value for $\mathbf{m}$ from equation 27;

4. Sample a new value for each α_l from equation 30;

5. Sample a new value for each b_{lk} from equation 33;

6. If the sample number is greater than the burn in period, store these new samples;

7. If we have not reached the maximum number of samples, return to 2, otherwise stop.

The length of burn-in period to take will depend on dataset. The easiest way to ensure that the chain(s) have converged is to start several and compute the $\hat{R}$ statistic (see, for example [8]) for each parameter of interest.

A.2 Classification

The full posterior is given by

$$p(\mathbf{Y}, \mathbf{M}, \boldsymbol{\alpha}|\mathbf{T}, \mathbf{C}, v, \lambda) = \prod_{k=1}^{K} p(\mathbf{M}_{\cdot k}|\mathbf{C}) \prod_{l=1}^{L} p(\alpha_l|v, \lambda) \prod_{n=1}^{N} p(t_{nl}|\mathbf{Y}^l_{n\cdot}) p(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l). \tag{34}$$

Taking the components one at a time, starting with the components of $\mathbf{Y}$, assuming that $t_{nl} = k$,

$$\begin{aligned} p(\mathbf{Y}^l_{n\cdot}|\dots) &\propto p(t_{nl}|\mathbf{Y}^l_{n\cdot}) p(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l) & (35)\\ &\propto \mathbf{1}(y^l_{nk} > y^l_{ni} \ \forall i \neq k) \mathcal{N}(\mathbf{Y}^l_{n\cdot}|\mathbf{M}_{n\cdot}, \alpha_l^{-1}\mathbf{I}_K), & (36) \end{aligned}$$

i.e., a Gaussian truncated such that the kth component is the largest. For $\mathbf{M}_{\cdot k}$

$$\begin{aligned} p(\mathbf{M}_{\cdot k}|\dots) &\propto p(\mathbf{M}_{\cdot k}|\mathbf{C}) \prod_{l=1}^{L} p(\mathbf{Y}^l_{\cdot k}|\mathbf{M}_{\cdot k}, \alpha_l) & (37)\\ &\propto \mathcal{N}(\mathbf{M}_{\cdot k}|\mathbf{0}, \mathbf{C}) \prod_{l=1}^{L} \mathcal{N}(\mathbf{Y}^l_{\cdot k}|\mathbf{M}_{\cdot k}, \alpha_l^{-1}\mathbf{I}_N) & (38)\\ &= \mathcal{N}(\mathbf{M}_{\cdot k}|\boldsymbol{\Sigma} \sum_{l=1}^{L} \alpha_l \mathbf{Y}^l_{\cdot k}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \mathbf{I}_N \sum_{l=1}^{L} \alpha_l \right)^{-1}. & (39) \end{aligned}$$

Finally, for the precisions,

$$\begin{aligned} p(\alpha_l|\dots) &\propto p(\alpha_l|v, \lambda) \prod_{k=1}^{K} p(\mathbf{Y}^l_{\cdot k}|\mathbf{M}_{\cdot k}, \alpha_l) & (40)\\ &\propto \mathcal{G}(\alpha_l|v, \lambda) \prod_{k=1}^{K} \mathcal{N}(\mathbf{Y}^l_{\cdot k}|\mathbf{M}_{\cdot k}, \alpha_l \mathbf{I}_N) & (41)\\ &= \mathcal{G}\left(v + \frac{NK}{2}, \lambda + \frac{1}{2} \sum_{k=1}^{K} (\mathbf{Y}^l_{\cdot k} - \mathbf{M}_{\cdot k})^T (\mathbf{Y}^l_{\cdot k} - \mathbf{M}_{\cdot k}) \right) & (42) \end{aligned}$$

The Gibbs sampler proceeds as follows

1. Initialise $\mathbf{m} \sim \mathcal{N}(\mathbf{m}|\mathbf{0}, \mathbf{C})$ and $\alpha_l \sim \mathcal{G}(\alpha_l|v, \lambda)$;

2. Sample a new value for each $\mathbf{Y}_{\cdot l}$ from equation 36;

3. Sample a new value for $\mathbf{m}$ from equation 39;

4. Sample a new value for each α_l from equation 42;

5. If the sample number is greater than the burn in period, store these new samples;

6. If we have not reached the maximum number of samples, return to 2, otherwise stop.

B A more efficient classification representation

The classification algorithm proposed requires a separate $\mathbf{Y}$ matrix to be maintained for each of the L experts. However, if we have reason to believe that each expert has the same noise precision $\alpha_l = 1$, we can greatly simplify this representation. Our likelihood, before performing the auxiliary variable trick, is as follows

$$p(\mathbf{T}_{n\cdot}|\mathbf{M}_{n\cdot}) = \prod_{k=1}^{K}\prod_{l=1}^{L}[\Phi^{kl}(\mathbf{M}_{n\cdot})]^{\delta(t_{nl}=k)} \tag{43}$$

where $\Phi^{kl}(\mathbf{M}_{n\cdot}) = p(t_{nl} = k|\mathbf{M}_{n\cdot})$. We notice that the product over the L experts can be replaced by raising the Gaussian to the power of a sum, i.e. we represent it as a multinomial distribution in terms of the sufficient statistic,

$$p(\mathbf{T}_{n\cdot}, \mathbf{Y}^1_{n\cdot}, \ldots, \mathbf{Y}^K_{n\cdot}|\mathbf{M}_{n\cdot}) = \prod_{k=1}^{K}[\mathcal{N}(\mathbf{Y}^k_{n\cdot}|\mathbf{M}_{n\cdot}, \mathbf{I}_K)]^{\sum_l \delta(t_{nl}=k)}. \tag{44}$$

So, rather than L matrices, we only need K, and the terms in the likelihood are raised to the power of the number of experts that assigned class k. Defining $\eta_{nk} = \sum_l \delta(t_{nl} = k)$, we can re-write this as

$$p(\mathbf{T}_{n\cdot}, \mathbf{Y}^1_{n\cdot}, \ldots, \mathbf{Y}^K_{n\cdot}|\mathbf{M}_{n\cdot}) = \prod_k N(\mathbf{Y}^k_{n\cdot}|\mathbf{M}_{n\cdot}, \eta^{-1}_{nk}\mathbf{I}_K). \tag{45}$$

For convenience, we have been slightly abusive with notation - $\mathbf{Y}^k_{n\cdot}$ only exists for k such that $\eta_{nk} > 0$. This will obviously be a different set for each n. We will hereafter refer to the complete set of $\mathbf{Y}^j_{n\cdot}$'s, over all n training points as $\mathbf{Y}$.

In the worst case, we will have to maintain a maximum of $\min(K, L)$ vectors for each data point, rather than L in the previous example. In all practical scenarios, the number will be far fewer than this as it is unlikely that the experts will all disagree on each training point. The corresponding conditional distributions are straightforward to extract and follow as

$$p(\mathbf{Y}^k_{n\cdot}|\mathbf{M}_{n\cdot}, t_{nl} = k) \propto \mathcal{N}(\mathbf{Y}^k_{n\cdot}|\mathbf{M}_{n\cdot}, \eta^{-1}_{nk}\mathbf{I}_K)\mathbf{1}(y^k_{nk} > y^k_{ni}\forall i \neq k), \tag{46}$$

i.e., a Gaussian truncated such that the kth value is largest and

$$\begin{aligned}
p(\mathbf{M}_{\cdot k}|\mathbf{Y}, \mathbf{C}, \mathcal{L}) &\propto \mathcal{N}(\mathbf{M}_{\cdot k}|\mathbf{0}, \mathbf{C})\prod_j\prod_n \mathcal{N}(y^j_{nk}|m_{nk}, \eta^{-1}_{nk})^{\mathbf{1}(\eta_{nj}>0)} && (47)\\
&= \mathcal{N}(\mathbf{M}_{\cdot k}|L\mathbf{\Sigma}_k\mathbf{z}_k\ ,\ \mathbf{\Sigma}_k) && (48)\\
\text{where } \mathbf{\Sigma}_k &= (\mathbf{C}^{-1} + L\mathbf{I}_N)^{-1} && (49)\\
\text{and } z_{nk} &= L^{-1}\sum_j \eta_{nj}y^j_{nk}. && (50)
\end{aligned}$$

We can now see that in the worst case we must store and update $\min(K, L)$ of the $\mathbf{Y}$ matrices. Generally, we will have to store and update fewer as it is unlikely that, in any practical application, all of the experts will disagree. It is worth noting that in these derivations, we have assumed that each expert has labeled each training point. It is straightforward to extend the algorithm to the more general case where each expert labels some subset of the total N examples.

C Predictive distributions for classification

In order to make predictions, we perform the following marginalisation

$$
\begin{aligned}
p(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{T}, \gamma, v, \lambda) &= \int\int \prod_{l=1}^{L} \int \mathbf{1}(y^l_{new,k} > y^l_{new,i}\ \forall i \neq k) p(\mathbf{Y}^l_{new\cdot}|\mathbf{M}_{new\cdot}, \alpha_l) && (51)\\
&\times\ p(\mathbf{M}_{new\cdot}\alpha_l|\mathbf{X}, \mathbf{T}, \gamma, v, \lambda) d\mathbf{Y}^1_{new\cdot}, \ldots, d\mathbf{Y}^L_{new\cdot} d\boldsymbol{\alpha} d\mathbf{M}_{new\cdot} && (52)\\
&= \int\int\int \mathbf{1}(y_{new,k} > y_{new,i}\ \forall i \neq k) \mathcal{N}\left(\mathbf{Y}_{new\cdot}|\mathbf{M}_{new\cdot}, \frac{1}{\sum_l \alpha_l}\mathbf{I}_K\right) && (53)\\
&\times\ p(\mathbf{M}_{new\cdot}\alpha_l|\mathbf{X}, \mathbf{T}, \gamma, v, \lambda) d\mathbf{Y}_{new\cdot} d\boldsymbol{\alpha} d\mathbf{M}_{new\cdot} && (54)\\
&= \int\int \mathrm{E}_{p(u)}\left[\prod_{j\neq k} \Phi\left(u + \left(\sum_l \alpha_l\right)^{1/2} (f_{new,k} - f_{new,j})\right)\right] && (55)\\
&\times\ p(\mathbf{M}_{new\cdot}\alpha_l|\mathbf{X}, \mathbf{T}, \gamma, v, \lambda) d\boldsymbol{\alpha} d\mathbf{M}_{new\cdot} && (56)\\
&\approx \frac{1}{N_{samps}} \sum_{s=1}^{N_{samps}} \mathrm{E}_{p(u)}\left[\prod_{j\neq k} \Phi\left(u + \left(\sum_l \alpha^s_l\right)^{1/2} (m^s_{new,k} - m^s_{new,j})\right)\right] && (57)
\end{aligned}
$$

where the supersctipt s corresponds to the sth sample from our Gibbs sampler and $f^s_{new,k}$ is drawn from a Gaussian with mean $\mathbf{M}^s_{\cdot k}\mathbf{C}^{-1}\mathbf{C}_{new}$ and variance $c_{new} - \mathbf{C}^T_{new}\mathbf{C}^{-1}\mathbf{C}_{new}$, $c_{new} = C(\mathbf{x}_{new}, \mathbf{x}_{new})$ and the ith element of the column vector $\mathbf{C}_{new}$ is $C(\mathbf{x}_i, \mathbf{x}_{new})$.